HOW-TO: make EPDnew database

René Dreos

May 23, 2018

Contents

1 Introduction

This dcument describes the steps to take and the software to used in order to generate a new release of the EPDnew database. The first part of the document covers the generation of the database and describes output file formats, how to check for possible errors and which files are generated. The second part describes how to transfer the final files to the desired locations and how to publish them. The last part provides a detailed description of the process itself, the software used and how to modify it. It is important to notice that this document does not cover the topics of how to generate a validated promoter collection.

2 Make a new release

This section of the document describes how to make a new release, the input files needed, how to check for possibe errors, where the final files are stored and how to move them to their final location on server01 in order to be accessed by the webserver.

2.1 Where

The right place to make a new EPDnew database release is the **working directory**:

```
ccg-serv02.vital-it.ch:/home/local/ccguser/epd_http/dbrelease/epdnew/
```

This is the working directory and also the place that contains the scripts, the input files and old database versions. This directory contains the following subdirectories:

./documents
: It contains useful documentation for generate a new database. It also contains a copy of this document.

./src
: It contains all the software needed to make a new database.

./errors
: It contains errors files that should be check before final approval of new releases.

./tmp
: Directory used by scripts to store temporary files. It can be ignored by the user.

./human
: This directory (and the twins for the other organisms) contains the relative database versions (old and new) and input files.

2.2 Files needed

Before making a new database release please make sure you have all the following files:

Promoter file:

An extended SGA file containing the genome coordinates of the validated promoters. This file should be independently generated. If you are interested in knowing how to generate a validated promoter collection please look at the `README.sh` file in the input directory for old EPDnew versions. An example can be found here:

```
/home/local/ccguser/epd_http/dbrelease/epdnew/human/new_promoters/012/README.sh
```

The following is an example of a promoter file for the human collection:

```
NC_000001.11   TSS   959256   -   806   ENST00000327044..NOC2L
NC_000001.11   TSS   960633   +   75    ENST00000338591..KLHL17
NC_000001.11   TSS   966482   +   80    ENSTR0000000002..PLEKHN1
NC_000001.11   TSS   976681   -   3     ENST00000479361..PERM1
NC_000001.11   TSS   1000097  -   462   ENST00000304952..HES4
```

The 3^{rd} field contains the validated TSS coordinate whereas the 5^{th} the number of samples in which the promoter has been validated. The 6^{th} field contains the transcript ID plus the gene name (it must be equal to "Gene Symbol" attribute in the annotation file) associated to that promoter separated by two dots [..]. Note also that the third line has a different type of transcript ID [***ENSTR****0000000002*]. This is a random ID characterised for having a fixed part (the letters) and a random part (the numbers) that has been generated during the promoter validation process. The fixed part marks this promoter as derived from a paired-end TSS mapping technique (in this case Rampage).

Annotaion file:

This is a tab delimited file, normally downloaded from BIOMART but can be generated from any other database (like UCSC table browser), with the foolowing attributes:

```
<Gene ID> <Transcript ID> <Gene Symbol> <RefSeq ID> <Description>
```

with:

- `<Gene ID>`: gene ID from one of the following databases: ENSEMBL (*H. sapiens*, *M. musculus*, *D. rerio*), FlyBase (*D. melanogaster*), BEEBASE (*A. mellifera*), WormBase (*C. elegans*), TAIR (*A. thaliana*), GRAMENE (*Z. mays*), PomBase (*S. pombe*), SGD (*S. cerevisiae*).
- `<Transcript ID>`: transcript ID from the same database as Gene ID
- `<Gene Symbol>`: "Gene Symbol" attribute must be equal to the gene names (6^{th} field) used in the Promoter file.
- `<Description>`: text description of the gene function

RNA file:
This is an ordered SGA file used to evaluate the promoter shape (focused, broad, etc...) and should be generated by concatenating all RNA samples (for example CAGE input files) that has been used to generate the validated promoter collection.

CpG island file:
This file is not mandatory as it is valid only for human. It is a BED file with the CpG islands boundaries annotated to the same assembly as the promoter file. It can be downloaded directly from the UCSC Genome Browser. Here is an example:

```
chr1    28735   29810   CpG:_116
chr1    135124  135563  CpG:_30
chr1    327790  328229  CpG:_29
chr1    437151  438164  CpG:_84
```

INI file:
This is a configuration file containing information (such as organism, genome assembly, input files phisical locations) that are needed by the scripts to make a new release. The next sub-section presents a detailed description of how to make it.

2.2.1 The INI configuration file

The INI configuration file contains the informations needed to make a database release. It has the following general structure:

```
PARAMETER_NAME = variable
```

The `PARAMETER_NAME` is fixed and can not be change by the user wheras the `variable` part can. Comment lines start with #.
The following are the accepted `PARAMETER_NAME` that must be present in each INI file:

`ORGANISM =`
It specifies the organism for which the database should be made and are equal to the final directories where databases are stored.

`ASSEMBLY =`
It specifies the assembly used to generate the promoter collection. It follows UCSC Genome Browser nomenclature. Accepted values are: hg39, mm9, danRer7, araTha1, sacCer3, etc...

`SGA_FILE =`
It specifies the **full path** to the "Promoter File".

`ENSEMBL_ANNOTATION =`
It specifies the **full path** to the "Annotation File".

`RNA_SGA =`
It specifies the **full path** to the "RNA File".

RNA_DIR =
: It specifies the **full path** to the data directory that store all sample SGA files used to generate the database.

CPG_FILE =
: It specifies the **full path** to the "CpG island file".

This is the INI file used to make the *H. sapiens* collection:

```
# Parameters used to generate the human promoter collection:

ORGANISM = human
ASSEMBLY = hg38

# The following is the SGA file with the new promoter
# collection. For an explanation how to make it please
# check file README.sh in the same directory.
SGA_FILE = /home/local/ccguser/epd_http/dbrelease/epdnew/ \\
  human/new_promoters/012/HsGencodePromotersValidated200Utr.sga

# Annotation file, contain data from BIOMART in the form
# <Gene ID> <Transcript ID> <Gene Symbol> <RefSeq ID>
# <Description>
ENSEMBL_ANNOTATION = /export/data/mga/hg38/gencode/ \\
  scripts/gencode28annotation.tsv

# This file is used to check the promoter type
# (broad or focused):
RNA_SGA = /home/local/ccguser/epd_http/dbrelease/ \\
  epdnew/human/new_promoters/011/allSamplesCageGt4.sga

# This is the directory containing all RNA samples:
RNA_DIR = /home/local/ccguser/epd_http/dbrelease/ \\
  epdnew/human/new_promoters/012/data/

# This is the BED file with CpG islands coordinates
CPG_FILE = /home/local/ccguser/epd_http/dbrelease/ \\
  epdnew/human/cpgIslandExt.hg38.bed

# to run:
# $ ./src/makeEPDnew.pl human.ini
```

2.3 Scripts to run

Once the input files are copied on `ccg-serv02` and the INI file is updated, starting the procedure is quite straightforward. Simply run this command in the working directory:

```
$ ./src/makeEPDnew.pl file.ini
```

This script automatically checks for the database version, errors and generate final files and folders.

2.4 Check for errors

The master script automatically checks for errors during execution. If a possible error is found it prints a warning message and highlights the relevant file in the `errors` directory. The user should inspect manually these files to check if the error is true. If this is the case, they should be fixed before upgrading the web-server. After the error has been fixed, the master script has to be run again.

2.5 Where is what

Once the master script has finished, all the relevant files have been moved to the directory:

`organism/version`

where `organism` is the organism specified in the INI file and `version` is the new release version number (this is automatically guessed by the master script). This directory contain the following files:

- `log_file.txt`: this is a text file containing all the relevant parameters used to generate this release
- `.dat` EMBL-like annotation file for each entry. A detailed description of this file can be found here: https://epd.vital-it.ch/current/usrman.php
- `.sga` file. this file contain information about the location of a promoter in the genome. It is a tab delimited file with the following fields:
  - Chromosome name with RefSeq identifier
  - Feature type. Here is always TSS (transcription start site)
  - Position in the chromosome (starting at base 1)
  - Strand (+ or -)
  - Count field (1)
  - Annotation field (Associated promoter ID)
- `.fps` file. Promoter collection in FPS format. This file format is used by our sister toolkit SSA (https://ccg.vital-it.ch/ssa/).
- `.bed` file. annotation file in bed format used to draw the EPDnew track on the UCSC Genome Browser. This is a representation of the sequence field (SE lines) of the DAT file. As such it starts at base -49 from the TSS (base 0) and end at base 10 (60 bp interval). BED file follows UCSC standards. The column are:

  `Chromosome RegionStart RegionEnd PromoterID Score Strand thickStart thickEnd`

with:

  - `Chromosome`, the chromosome with UCSC nomenclature
  - `RegionStart`, the start of the region (base -49 or + 10 from the TSS)
  - `RegionEnd`, the end of the region (base -49 or + 10 from the TSS)
  - `PromoterID`, EPDnew promoter ID
  - `Score`, required by UCSC, always 900
  - `strand`, + or -
  - `thickStart`, start of the thich region (TSS or base +10)
  - `thickEnd`, end of the thich region (TSS or base +10)

  note that BED files start chromosomes at base 0, whereas SGA chromosomes start at base 1. Moreover the bed file is chromosome oriented, so the TSS can be found at the start or at the end of the thick region depending on the strand.

- `.bb` file. Promoter collection in BigBed format, derived from the BED file.
- `.idx` file. Files with indeces used by in-house software to rapidly access the database [mostly outdated].

If a new human collection is generated for assebly hg38, the master script also liftOvers the promoter collection to hg19.

- `db` directory: it contains files used to upgrade the MySQL database. They share the same name between different organisms:
  - `promoter_coordinate.txt` contains information on promoter location in the genome, organism and promoter type. Columns are the following:
    * EPDnew promoter ID
    * chrmosome RefSeq ID
    * TSS position as defined by EPDnew
    * strand: + or -
    * scientific name of the organism
    * type: either “single”, “multiple” or “region”
  - `promoter_samples_expression.txt` contains average expression levels for each promoter with the following columns:
    * EPDnew promoter ID
    * number of samples in which the promoter is active
    * average expression level (evaluated as the average number of CAGE tags that map in a window of 100 bp arount annotated TSS)
  - `promoter_expression.txt` contains information on expression levels of each promoter in each sample used during EPDnew validation process. Columns are the following:
    * EPDnew promoter ID

    - expression level (evaluated as the total number of CAGE tags that map in a 100bp window around the TSS)
    - sample-specific TSS position relative to the annotated TSS
    - sample name
- `promoter_sequence.txt` contains sequence information for each promoter. Columns are the following:
    - EPDnew promoter ID
    - short sequence segment corresponding to the -49 to +10 region of the promoter
- `promoter_ensembl.txt` links EPDnew IDs with ENSEMBL Gene IDs. Columns are the following:
    - EPDnew promoter ID
    - ENSEMBL Gene ID
- `cross_references.txt` links ENSEMBL Gene IDs with external databases ID:
    - ENSEMBL Gene ID
    - Gene Name
    - RefSeq ID
    - Gene Description
- `gene_description.txt` stores information on gene name and description. Columns are the following:
    - EPDnew promoter ID
    - Gene Name
    - Gene Description
- `promoter_motifs.txt` is a boolean file describing the presence or absence of core promoter elements. Columns are the following:
    - EPDnew promoter ID
    - TATA-box presence at position -28 (+- 3 bp)
    - Initiator presence at the TSS
    - CCAAT-box presence in the region -200 to -50 from the TSS
    - GC-box presence in the region -200 to -50 from the TSS
- `promoter_ucsc.txt` contains information about the promoter location in UCSC style format. Columns are the following:
    - EPDnew promoter ID
    - UCSC assembly name (e.g. hg19, hg38, ...)
    - UCSC chrmosome name (e.g. chr1, chr2, ...)
    - strand (+ or -)
    - position in the genome (base 0 start)

3 Publish the new release

Publishing a new release require a few manual steps such as files transfer to the server, updating web pages and the UCSC hubs. Here is a detailed description of what to do to make your new release publicly accessible on–line.

3.1 Move files to final destination

Once all the errors have been checked and corrected, the final files can be moved to their final destination on `ccg-serv01` (the backup server) and `ccg-serv04` (the web server) using an `rsync` command:

```
cd organism
rsync -avz version ccg-serv01:/home/local/ftp/epdnew/organism/
rsync -avz version ccg-serv04:/home/local/ftp/epdnew/organism/
```

where `organism` is the organism specified in the INI file and `version` is the new release version number. Moreover, in the destination directory on `ccg-serv01` and `ccg-serv04`, the symbolic links in the `current` directory have to be manually upgraded to the new version. For example with a command like this:

```
cd /home/local/ftp/epdnew/organism/current/
for I in bed dat fps idx sga; do
   rm Hs_EPDnew.$I;
   ln -s ../006/Hs_EPDnew_006_hg38.$I Hs_EPDnew.$I;
done
```

3.2 Update MySQL database

CAUTION: do this step only after the `ccg-serv04` web-server has been upgraded with the new promoter database files (normally the day after they have been transfered to `ccg-serv01`). To be sure, check if the new version is present in the ftp directory on `ccg-serv04`:

```
/home/local/ftp/epdnew/organism/version/
```

The web server "Search" and "Viewer" pages get their information from species-specific `MySQL` databases. It is important to upload them with the new promoter collection for it to be accessible on the web. Here is how to do it:

1. On `ccg-serv04` (the web server) as `ccguser` go to the directory

   ```
   /home/local/ccguser/epd_http/data/mysql/
   ```

2. Open the relevant `.sql` file (for example, if uploading the human database open the file `make_human_database.sql`). Change the version number in all `LOAD` commands to the new version so that the script can load the newest files. Save and close the file.

3. Run the following command:

   ```
   mysql < make_human_database.sql
   ```

   Carefully check that there are no errors. If so exam and solve them. In the mean time revert to the previous stable version (change the `.sql` file back to the original version and run the command).

3.3 Update the UCSC Hub

EPDnew viewer page makes use of 3 UCSC Genome Browser Hubs. Two (`epdHub` and `epdHub2`) are for assemblies supported by UCSC (*H. sapiens*, *M. musculus*, *D. rerio*, *D. melanogaster*, *C. elegans*, *S. cerevisiae*), the other one (`epdHubCustomSpecies`) is for species not supported (*S. pombe*, *A. thaliana*, *Z. mays*, *A. mellifera*). `epdHub` and `epdHub2` differ in a few settings and are used for different purposes: `epdHub` is used for links to UCSC Genome Browser whereas `epdHub2` to download the promoter picture. Here you can find a short guide on Track Hubs:
https://genome.ucsc.edu/goldenpath/help/hubQuickStartAssembly.html
In general a Hub is structured as a directory tree containing text files and data that are interpreted by UCSC and used to plot a customised version of the browser. The root directory of all hubs can be found here:

`/home/local/ccguser/epd_http/htdocs/ucsc/`

and within each Hub directory there are several subdirectories for each assembly (for example `hg38`, `mm9`). Whithin each assembly directory there is a configuration file (`trackEpd.txt`) that need to be updated with the new release. Please note that for the species for which we support several assemblies (eg. *H. sapiens*), you have to update all of them.

The procedure to modify the configuration file is the same for all assemblies:

1. go to the corresponding assembly directory and open the `trackEpd.txt` configuration file
2. find the block corresponding to EPDnew track (normally at the end of the file)
3. duplicate the EPDnew block
4. change `track` name, `shortLabel`, `longLabel`, `bigDataUrl` fields to the new version for one of the copies
5. hide the old version changing the `visibility` setting from `full` to `hide`

Once the UCSC hubs are updated, delete the stored promoter images (a script will automatically download the new images when needed) that are stored here:

`/home/local/ccguser/epd_http/htdocs/epdnew/gif/organism/`

3.4 Update database home page and assembly pipeline description

EPDnew species-specific databases have their own hompage. When a new release is published the home page and assembly pipeline description page need to be updated. The homepage can be found here:

`/home/local/ccguser/epd_http/htdocs/organism/organism_database.php`

with `organism` one of the 10 supported species (for example: `human`, `mouse`, etc.). The assembly pipeline description page instead can be found here (this is the human page):

`/home/local/ccguser/epd_http/htdocs/epdnew/documents/Hs_epdnew_006_pipeline.php`

the pages for the other organisms can be found in the same directory. Note that each page is version-specific and old version pages are kept as reference.

3.5 Add the new promoter collection to the MGA database

On `ccg-serv01`:

1. Copy the `.sga` and `.fps` files to the assembly–specific MGA epd directory:

   `/export/data/mga/assembly/epd/`

2. Add the new release to the `epd.txt` and `epd.html` files that are located in the same directory

3.6 Write a news item

Once everything is done you can add a news item about the new release using the following web–page: https://epd.vital-it.ch/addNews.php. Note that you have to do this on a machine with the following ID addresses: `128.178.198.221`, `128.178.198.224`, `128.178.198.225` and `128.178.198.226`.

4 Software description

This section briefly describes the main script that generate EPDnew database. It can be found in the directory `src/`.

`makeEPDnew.pl <file.ini>`

This is the main script. It sequentially calls all the scripts needed to generate a new database. The only required parameter is the path to the INI configutration file. The script is heavily commented and should be quite easy to read and understand what it is doing. The procedure is the following:

1. Parse ini file.
2. Get version number.
3. Add EPD promoters to collection (only if the same gene has not been already validated).
4. Rank promoters belonging to the same gene by their expression level (first promoter [_1] is the strongest).
5. Make the DAT file and check that promoter names are not duplicated in different chromosomes or strands.
6. Add external IDs to DAT file.
7. Make SGA and FPS files.
8. Check promoter initiation type (single, broad, region).

9. Make BED file.
10. Make BigBed file.
11. Generate hg19 version (only for *H. sapiens*).
12. Make files for MySQL database.
13. Make IDX file.
14. Move file to final destination in the working directory.

5 Info about this document

This document has been written using LaTeX. The raw document can be found here:

`sv8-225.epfl.ch:/home/rdreos/Work/epd/epd_new/documents`